DOI: 10.5604/01.3001.0010.7371

WPŁYW FUNKCJI OKNA NA SKUTECZNOŚĆ IDENTYFIKACJI STANU EMOCJONALNEGO MÓWCY

Paweł Powroźnik, Dariusz Czerwiński
Politechnika Lubelska, Instytut Informatyki

Streszczenie. *Artykuł prezentuje wpływ doboru funkcji okna wykorzystywanej w procesie obliczania spektrogramu, na skuteczność identyfikacji stanu emocjonalnego mówcy posługującego się mową polską. W badaniach wykorzystano następujące funkcje okna: Hamminga, Gaussa, Dolpha–Czebyszewa, Blackmana, Nuttalla, Blackmana-Harrisa. Ponadto został przedstawiony sposób przetwarzania spektrogramu przez sztuczną sieć neuronową (SSN), odpowiedzialną za identyfikację stanu emocjonalnego mówcy. Otrzymane wyniki pozwoliły na ocenę skuteczności rozpoznawania stanu emocjonalnego za pomocą SSN. Średnia skuteczność wahała się od około 70% do ponad 87%.*

Słowa kluczowe: okna, sztuczne sieci neuronowe, identyfikacja polskiej mowy emocjonalnej

THE IMPACT OF WINDOW FUNCTION ON IDENTIFICATION OF SPEAKER EMOTIONAL STATE

Abstract. *The article presents the impact of window function used for preparing the spectrogram, on Polish emotional speech identification.. In conducted researches the following window functions were used: Hamming, Gauss, Dolph–Chebyshev, Blackman, Nuttall, Blackman-Harris. The spectrogram processing method by artificial neural network (ANN) was also described in this article. Obtained results allowed to assess the effectiveness of identification process with the use of ANN. The average efficiency ranged from 70 % to more than 87%.*

Keywords: window function, artificial neural networks, Polish emotional speech recognition

Wstęp

Identyfikacja stanu emocjonalnego mówcy przez systemy informatyczne jest zadaniem skomplikowanym, jednak coraz bardziej pożądanym w dzisiejszym świecie. Według badań człowiek jest w stanie poprawnie zidentyfikować stan emocjonalny nieznanej mu osoby w około 60% przypadków [22]. Zatem wyzwanie jakie stoi przed systemami informatycznymi do automatycznej detekcji jest duże.

Oczywistym jest, iż informacje przekazywane poprzez intonację głosu zmieniają charakter wypowiadanego tekstu. Te same zdania nacechowane różnymi emocjami mogą mieć zgoła odmienne znaczenia. Z punktu widzenia systemów informatycznych służących do rozpoznawania mowy idealną sytuację stanowi wypowiedź nie nacechowana emocjonalnie ze względu na brak transferu dodatkowych informacji [20].

Innymi słowy emocje występujące w sygnale mowy mogą znacząco pogorszyć dokładność systemów automatycznego jej rozpoznawania [10].

Największy problem w rozpoznawaniu mowy emocjonalnej stanowi mnogość stanów emocjonalnych. Opracowanie systemu poprawnie identyfikującego większość emocji nie jest zagadnieniem trywialnym, dlatego też w pracach badawczych najczęściej rozważane są następujące stany emocjonalne: radość, smutek, strach, znudzenie, złość oraz stan neutralny [10, 21].

W niniejszym artykule został przedstawiony wpływ funkcji okna, wykorzystywanej w procesie tworzenia spektrogramu, na skuteczność identyfikacji stanu emocjonalnego mówcy przez sztuczną sieć neuronową (SSN). W przeprowadzonych badaniach wykorzystane zostały następujące okna czasowe: Hamminga, Gaussa, Dolph-Czebyszewa, Blackmana, Nuttala i Blackmana-Harrisa. Ponadto został przedstawiony nowatorski sposób przetwarzania spektrogramu, będącego wektorem wejściowym dla SSN.

Artykuł został podzielony na cztery części. Pierwsza z nich charakteryzuje zakres badań. W drugiej została przedstawiona wykorzystywana baza nagrań mowy emocjonalnej oraz omówiona została analiza widmowa sygnału mowy. Trzecia część traktuje o zastosowanych metodach badawczych. Ponadto zostały scharakteryzowane struktury wykorzystanych sieci neuronowych. W części tej zaprezentowano również otrzymane wyniki. Na końcu artykułu zostały nakreślone możliwe kierunki rozwoju badań oraz propozycje poprawy wykorzystanych metod.

Należy zauważyć, iż opracowana metoda identyfikacji stanu emocjonalnego mówcy daje zdecydowanie lepsze wyniki od poprzednio uzyskiwanych podczas pracy nad tym samym zbiorem nagrań, gdzie średnia skuteczność identyfikacji stanu emocjonalnego wynosiła około 50% [19].

1. Krótka charakterystyka analizowanego zagadnienia

Najbardziej złożonym zagadnieniem w identyfikacji stanu emocjonalnego mówcy jest mnogość emocji, które należy rozpoznać. W literaturze pojawiają się wzmianki o sposobach przetwarzania sygnału polskiej mowy emocjonalnej oparte na maszynie wektorów wspierających [9] czy algorytmie k-NN [10] jednak wyniki otrzymywane przy ich wykorzystaniu ciągle nie są w pełni zadowalające. Niniejsze badania skupiały się na sześciu najpopularniejszych stanach: strachu, smutku, znudzeniu, radości, złości i stanem neutralnym.

Metody czasowo – częstotliwościowe są najczęściej wykorzystywanymi narzędziami w przetwarzaniu sygnałów mowy [17]. Mogą one być podzielone na dwie zasadnicze grupy: reprezentacja czas – skala i czas – częstotliwość. Metody te umożliwiają estymację widma sygnału w krótkim i skończonym przedziale czasowym, bazując na fragmentach sygnału pozyskiwanego przez przesuwne okno czasowe [17].

Szczególną rolę w przetwarzaniu sygnału mowy pełni krótkoczasowa transformata Fouriera (STFT) i związany z nią spektrogram. Bardzo ważną rolę w procesie powstawania spektrogramu odgrywają funkcje okna czasowego. Wpływ tych funkcji na skuteczność rozpoznawania stanu emocjonalnego mówcy jest głównym przedmiotem badań opisanym w niniejszym artykule. Jako klasyfikatorem posłużono się sztuczną siecią neuronową. W skład wektora uczącego wchodził w 50% zbiór posiadanych nagrań emocjonalnej mowy polskiej. Jako zbiór testowy posłużyło pozostałe 50% nagrań. Ze względu na ograniczoną ilość danych, nagrania uczące i testowe pochodziły od tych samych mówców jednakże została zachowana niezależność od tekstu (trzy pierwsze zdania z bazy danych posłużyły jako wektory uczące, trzy kolejne jako testowe).

2. Analiza widmowa sygnału mowy

Aby dokonać analizy sygnału mowy należy w pierwszej kolejności zamienić sygnał analogowy na sygnał dyskretny. Odbywa się to poprzez proces próbkowania z określoną częstotliwością. Sygnał cyfrowy może w dalszym etapie posłużyć do przeprowadzenia analizy.

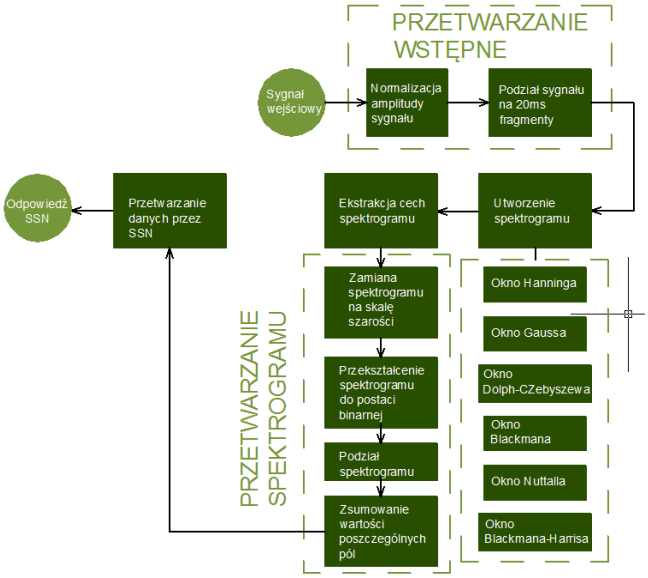
Zastosowanie metod wykorzystujących konwencjonalne podejście np. wyznaczanie widma długoterminowego

w przypadku sygnałów mowy nie daje zadowalających wyników ze względu na specyfikę tych sygnałów. Z tego powodu najczęściej stosowaną widmową metodą analizy sygnałów mowy jest metoda wykorzystująca wyznaczanie chwilowego widma sygnału i przedstawienie jego ewolucji w czasie na tzw. spektrogramie [25].

Największe wyzwaniem w przeprowadzonych badaniach stanowił proces przygotowania i przetwarzania danych. Został on podzielony na kilka etapów.

Pierwszy obejmował wstępne przetwarzanie sygnału wejściowego. Jego pierwszym etapem była normalizacja wartości danych wejściowych do przedziału [-1,1]. W zagadnieniach związanych z przetwarzaniem mowy powszechnym zjawiskiem jest podział sygnału wejściowego na ramki o stałej długości (ramkowanie) [6]. W niniejszych badaniach posłużono się ramkami o długości 20ms każda. Jednym z problemów związanym z próbkowaniem i ramkowaniem jest możliwość wystąpienia zjawiska "wycieku danych". Jest to spowodowane nieciągłą zmianą sygnału na końcach przedziału próbkowania. W celu ograniczania występowania tego zjawiska należałoby zastosować metodę zwaną okienkowaniem [3, 5, 12, 15] i polegającą na przemnożeniu ciągu wejściowego przez funkcję okna powodującą redukcję amplitudy prążków widma przede wszystkim na końcach ramek, minimalizując wpływ składowych wysokoczęstotliwościowych będących powodem przecieków.

Kolejny etap skupiał się wokół transformacji wstępnie przetworzonego sygnału do postaci spektrogramu. W tym etapie wykorzystywane były funkcje okna, których dobór stanowił jedno z zagadnień opisywanych w niniejszym artykule. W przeprowadzonych badaniach użyte zostały okna o długości 128 próbek, z kolei nakładkowanie wynosiło 50%. Ostatnim krokiem była ekstrakcja cech stanowiących wektor wejściowy dla SSN. Całość procesu przetwarzania sygnału została przedstawiona na rys. 1.



Rys. 1. Proces przetwarzania sygnału

2.1. Opis wykorzystanej bazy danych

W badaniach związanych z identyfikacją mowy emocjonalnej bardzo często stosowana jest Berlińska Baza Danych Mowy Emocjonalnej (BES) [1] przygotowana przez zawodowych aktorów i zawierająca nagrania w siedmiu stanach emocjonalnych to jest: strach, złość, znudzenie, radość, smutek, odraza oraz stan neutralny. Jednakże, w przypadku polskiej mowy emocjonalnej większość badaczy skupia się wokół bazy danych opracowanej przez Zakład Elektroniki Medycznej Politechniki Łódzkiej [5]. Baza ta została przygotowana przez ośmioro aktorów: czterech mężczyzn i cztery kobiety. Zebrane nagrania występują w sześciu stanach emocjonalnych analizowanych w niniejszym artykule. Cały zbiór składa się z 240 nagrań w formacie '.wav' próbkowanych z częstotliwością 44,1 kHz. Baza ta zawiera następujące nagrania: „Od dziś przestaję się golić”, „Janek był dzisiaj u fryzjera”, „Oni kupili dzisiaj nowy samochód”, „Ta lampa dzisiaj jest na biurku” oraz „Jego dziewczyna przylatuje dzisiaj samolotem".

2.2. Okna czasowe

Okna czasowe są funkcjami spełniającymi następujące warunki [4]:

- są symetryczne względem środka przedziału,
- są niezerowe w skończonym przedziale czasu,
- osiągają maksimum w środku przedziału.

Okienkowanie polega na wykonaniu operacji splotu sygnału wejściowego oraz funkcji okna w osi czasu. Wynikiem powyższej operacji jest zmiana amplitudy sygnału w funkcji okna.

Okno Hamminga

Okno czasowe jest funkcją opisującą sposób pozyskiwania próbek z analizowanego sygnału [26]. Przy założeniu, iż dany jest pewien sygnał $s(n)$, w skończonym przedziale czasu, wówczas wynikiem obserwacji takiego impulsu w oknie będzie funkcja $g(n)$ opisana następującym wzorem:

$$g(n)=s(n)w(n), \qquad n\in(-\infty,+\infty), \tag{1}$$

gdzie $w(n)$ jest wspomnianą funkcją okna [26].

Szczególny przykład okna czasowego stanowi zaproponowane przez R. W. Hamminga okno Hamminga. Zostało ono opracowane aby minimalizować wartość maksymalną najbliższego płatka bocznego i charakteryzowane jest następującym wzorem [8]:

$$w(n)=\alpha-\beta\cos\left(\frac{2\pi n}{N-1}\right), \tag{2}$$

gdzie: $\alpha=0{,}54$, $\beta=1-\alpha=0{,}46$, N – liczba próbek sygnału.

Okno Gaussa

Okno czasowe Gaussa zdefiniowane jest następującym wzorem [8]:

$$w(n)=e^{-\frac{1}{2}\left(\frac{n-(N-1)/2}{\sigma(N-1)/2}\right)^2}, \tag{3}$$

gdzie: N – liczba próbek sygnału, $\sigma\le 0{,}5$

Okno to posiada dwie podstawowe zalety w odniesieniu do transformaty Fouriera. Kształt funkcji Gaussa zbliżony jest do paraboli, zatem może być niemal dokładnie wykorzystany w kwadratowej interpolacji estymacji częstotliwościowej. Po drugie w wyniku transformacji Fouriera funkcji Gaussa otrzymujemy również funkcję Gaussa zatem jest to funkcja własna transformacji.

Okno Dolpha-Czebyszewa

Okno Dolpha-Czebyszewa jest definiowane następująco [16]:

$$w(n)=w_0\left(n-\frac{N-1}{2}\right), \tag{4}$$

gdzie:

$$w_0(n)=\frac{1}{N}\sum_{k=0}^{N-1}W_0(k)\;e^{\frac{i2\pi kn}{N}} \qquad n\in\left\langle-\frac{N}{2},\frac{N}{2}\right\rangle, \tag{5}$$

$$W_0(k)=\frac{\cos\left\langle N\cos^{-1}\left[\beta\cos\left(\frac{\pi k}{N}\right)\right]\right\rangle}{\cosh\left[N\cosh^{-1}(\beta)\right]} \tag{6}$$

gdzie:

$$\beta=\cosh\left[\frac{1}{N}\cosh^{-1}\left(10^{\alpha}\right)\right], \tag{7}$$

gdzie: α jest parametrem określonym za pomocą norm Czebyszewa i określany jest jako logarytm stosunku wysokości maksimum głównego do bocznych [5].

Okno Blackmana

Opierając się na opracowaniu [7] okno Blackmana jest zdefiniowane w następujący sposób:

$$w(n) = a_0 - a_1 \cos\left(\frac{2\pi n}{N-1}\right) + a_2 \cos\left(\frac{4\pi n}{N-1}\right), \quad (8)$$

gdzie: $a_0 = \frac{1-\alpha}{2}$, $a_1 = \frac{1}{2}$, $a_2 = \frac{\alpha}{2}$, $\alpha = 0{,}16$.

Okno Nuttalla

Matematycznie okno Nuttala opiera się na następującej definicji [8]:

$$w(n) = a_0 - a_1 \cos\left(\frac{2\pi n}{N-1}\right) + a_2 \cos\left(\frac{4\pi n}{N-1}\right) - a_3 \cos\left(\frac{6\pi n}{N-1}\right), \quad (9)$$

gdzie: $a_0 = 0{,}355768$, $a_1 = 0{,}487396$, $a_2 = 0{,}144232$, $a_3 = 0{,}012604$.

Ponadto jeśli założymy, iż dana jest dowolna liczba rzeczywista *n*. Wówczas funkcja Nuttalla i jej pierwsza pochodna są zawsze ciągłe. Oznacza to, iż funkcja dąży do 0 dla *n* = 0.

Okno Blackmana-Harrisa

Okno to jest uogólnieniem okna Hamminga. Otrzymywane jest poprzez większe w przesunięcie funkcji *sinc* w celu minimalizacji prążków bocznych. Okno Blackama-Harrisa jest definiowane następująco [16]:

$$w(n) = a_0 - a_1 \cos\left(\frac{2\pi n}{N-1}\right) + a_2 \cos\left(\frac{4\pi n}{N-1}\right) - a_3 \cos\left(\frac{6\pi n}{N-1}\right), \quad (10)$$

gdzie: a_0 = 0,3635819, a_1 = 0,4891775, a_2 = 0,1365995, a_3 = 0,0106411.

2.3. Krótkoczasowa transformata Fouriera

Krótkoczasowa transformata Fouriera (STFT) pełni znaczącą rolę w analizie sygnału mowy podobnie jak metody spektrograficzne. Oba sposoby przetwarzania dźwięku są zaliczane do reprezentacji sygnału mowy w przestrzeni czas – częstotliwość [11]. Ciągła STFT może być interpretowana jako szczególny przypadek przekształcenia Gabora [26]. Dla ciągłego sygnału $x(t)$ transformacja jest zdefiniowana następująco [13]:

$$STFT_x^F(t,f) = e^{-j2\pi ft} \int_{-\infty}^{+\infty} X(\theta) W^*(\theta - f) e^{j2\pi\theta t} d\theta \quad (11)$$

Z kolei w dziedzinie czasu transformacja jest zdefiniowana następująco:

$$STFT_x^T(t,f) = \int_{-\infty}^{+\infty} x(\tau) w^*(\tau - t) e^{j2\pi f\tau} d\tau, \quad (12)$$

gdzie: $w(t)$ jest funkcją okna o widmie Fouriera $W(\theta)$, $X(\theta)$ – widmo analizowanego sygnału, znak "*" – oznacza sprzężenie zespolone.

Równanie (12) polega na wykonaniu przekształcenia Fouriera na następujących po sobie fragmentach sygnału wejściowego pozyskiwanych za pomocą okna $w(t)$. Krótkoczasowa transformata Fouriera w dziedzinie częstotliwości jest równoważna [4]:

1) odwrotnemu przekształceniu Fouriera obliczonemu dla fragmentu widma sygnału $X(\theta)$ pozyskanemu przez przesuwane w dziedzinie częstotliwości okno $W(\theta - f)$.
2) przemieszczeniu w częstotliwości sygnału pozyskanego z 1. do częstotliwości zerowej. Przekształcenie to wykonywane jest poprzez przemnożenie sygnału z 1. przez $e^{j2\pi f\tau}$.

Dla sygnałów cyfrowych szczególne znaczenie przyjmuje następująca postać powyższego równania [13]:

$$STFT(nk) = \sum_{m=-\infty}^{+\infty} x(m) w^*(n-m) e^{-j\left(\frac{2\pi}{N}k\right)m} \quad (13)$$

Jeżeli rozważymy funkcję okna o *N* niezerowych i rzeczywistych próbkach, wówczas powyższe równanie przyjmie następującą postać [13]:

$$STFT(nk) = \sum_{m=0}^{N} w(m) x(n-m) e^{-j\left(\frac{2\pi}{N}k\right)m} \quad (14)$$

gdzie: n = 0, *N*, 2*N*, ..., *M*-*N*; k = 0, 1, 2, ..., *N*-1, *M* – liczba analizowanych próbek.

Związek pomiędzy krótkoczasową transformatą Fouriera, a spektrogramem definiuje następujące równanie [13]:

$$S(nk) = |STFT(nk)|^2 \quad (15)$$

Obliczenia są wykonywane w oparciu o metodę przesuwnego okna w dziedzinie czasu. Krótkoczasowa transformata Fouriera jest obliczana dla każdego zbioru próbek ograniczonych przez okno czasowe.

Dobór rozdzielczości w obu dziedzinach ma zasadniczy wpływ na jakość spektrogramu. Szerokie okno znacząco zwiększa rozdzielczość w dziedzinie częstotliwości, negatywnie wpływając na jakość w dziedzinie czasu. Uzasadnienia takiego zjawiska należy szukać w zależnościach czasowo – częstotliwościowych funkcji okna. Uzyskanie wysokiej rozdzielczości czasowej wiąże się z koniecznością użycia wąskiego okna. Mała liczba próbek sprawia, iż obliczenia transformaty Fouriera wykonywane z krokiem $\Delta f = \frac{f_p}{N}$, f_p – częstotliwość próbkowania, będą realizowane ze znacznym przyrostem częstotliwości [2]. Dodatkowo w powyższym przypadku duża szerokość listka głównego na charakterystyce częstotliwościowej spowoduje pojawienie się efektu rozmycia [2]. Rozwiązaniem problemu związanego z doborem szerokości okna jest metoda nakładkowania, która znacząco poprawia jakość spektrogramu [14].

3. Przeprowadzone badania

Głównym celem przeprowadzonych badań było sprawdzenie wpływu funkcji okna, w procesie opracowywania spektrogramu, na skuteczność identyfikacji stanu emocjonalnego mówcy posługującego się mową polską. W badaniach została wykorzystana baza nagrań mowy emocjonalnej przygotowana przez Zakład Elektroniki Medycznej Politechniki Łódzkiej [5].

3.1. Dobór parametrów spektrogramu

Uzyskanie spektrogramu umożliwiającego efektywną identyfikację sygnału emocjonalnej mowy polskiej wiąże się z odpowiednim doborem takich parametrów jak: szerokość okna, funkcja okna czy rozdzielczość w dziedzinie czasu. Najwyższa rozdzielczość w dziedzinie czasu może być uzyskana przy zastosowaniu nakładkowania wynoszącego *N*-1 próbek, jednak jak łatwo zauważyć przesuwanie okna w każdym kroku tylko o jedną próbkę wiąże się ze znaczącym wzrostem obliczeń. Dlatego też w przeprowadzonych badaniach wykorzystano nakładkowanie wynoszące 50% długości okna. Właściwy dobór długości okna jest zagadnieniem nieco bardziej złożonym.

Najwyższa efektywność osiągana jest w sytuacji gdy stosunek średniokwadratowej długości częstotliwościowej (A) do czasowej (B) był równy stosunkowi przyrostu częstotliwości do czasu, w którym dany przyrost miał miejsce [2]:

$$\frac{A}{B} = \frac{\Delta f}{\Delta t}, \quad (16)$$

gdzie:

$$A = \sqrt{\frac{1}{E}\int_{-\infty}^{+\infty} f^2 \left|W(f)^2\right| df}, \quad (17)$$

$$B = \sqrt{\frac{1}{E}\int_{-\infty}^{+\infty} t^2 \left|w(t)^2\right| dt}, \quad (18)$$

$$E = \sqrt{\frac{1}{E}\int_{-\infty}^{+\infty} \left|w(t)^2\right| dt} = \sqrt{\frac{1}{E}\int_{-\infty}^{+\infty} \left|W(f)^2\right| df} \quad (19)$$

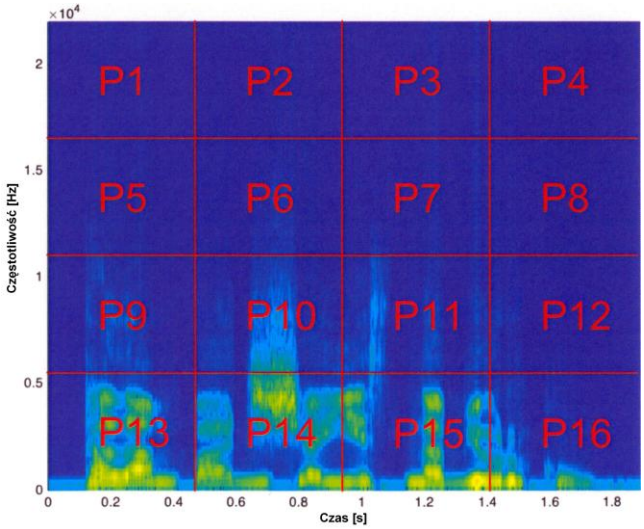
Problem doboru rodzaju funkcji okna jak i jej parametrów powinien stanowić swoisty kompromis pomiędzy jakością sygnału uzyskiwaną na wyjściu, a czasem niezbędnym do wykonania obliczeń. Należy również zauważyć, iż sam dobór funkcji okna jest pewnym kompromisem pomiędzy szerokością listka głównego, poziomem pierwszego listka bocznego oraz szybkością zmian poziomów listków bocznych wraz ze wzrostem częstotliwości. A zetem jest to kompromis pomiędzy dokładnością wartości amplitudy oraz częstotliwością.

3.2. Ekstrakcja cech spektrogramu

Głównym elementem etapu ekstrakcji cech ze spektrogramu było utworzenie zestawu danych stanowiących wektor wejściowy dla SSN. Proces ekstrakcji cech przebiegał następująco:

1) Przedstawienie spektrogramu w skali odcieni szarości (0-255).
2) Przekształcenie otrzymanego spektrogramu w postać binarną. Wartości poniżej progu zyskały wartość 0, powyżej – 1. Przeprowadzony został szereg eksperymentów mający na celu wyznaczenie najlepszej wartości progu. Zbadany został zakres od 100 do 200. Najlepsze wyniki zostały osiągnięte gdy wartość progu wynosiła 155.
3) Obraz uzyskany poprzez zamianę spektrogramu do obrazu binarnego został podzielony odpowiednio na 9, 16 i 25 fragmentów. Dla każdego z podziałów przeprowadzone zostały badania. Przykład podziału spektrogramu został przedstawiony na rysunku 2.

Wartości w poszczególnych obszarach zostały zsumowane stając się wektorem wejściowym dla SSN.



Rys. 2. Przykład podziału spektrogramu (dla lepszej czytelności został przedstawiony spektrogram w skali barwnej)

3.3. Zastosowanie sztucznych sieci neuronowych do identyfikacji emocji

W niniejszych badaniach zostały wykorzystane SSN udostępniane za pośrednictwem pakietu Neural Network Toolbox programu Matlab 2015b. W zależności od podziału spektrogramu zostały wykorzystane trzy różne SSN.

W przypadku podziału spektrogramu na 9 obszarów użyta sztuczna sieć neuronowa składała się z 4 warstw. Warstwę wejściową stanowiło 10 neuronów (9 ze spektrogramu i płeć mówcy). SSN miała dwie warstwy ukryte po 20 neuronów każda oraz 6 neuronów w warstwie wyjściowej odpowiadające 6 identyfikowanym stanom emocjonalnym.

W przypadku podziału spektrogramu na 16 fragmentów, warstwa wejściowa SSN składała się z 17 neuronów. Warstwy ukryte zawierały po 34 sztuczne komórki nerwowe każda. Z kolei warstwa wyjściowa zbudowana została z 6 neuronów.

Ostanie badania zostały przeprowadzone dla podziału spektrogramu na 25 obszarów. Warstwę wejściową stanowiło 26 neuronów. Warstwy ukryte zostały zbudowane odpowiednio z 20 i 10 neuronów. Natomiast podobnie jak w poprzednich przypadkach w warstwie wyjściowej znajdowało się 6 neuronów.

We wszystkich badaniach jako funkcja aktywacji neuronów została użyta funkcja tangens hiperboliczny. SSN były uczone za pomocą algorytmu wstecznej propagacji błędów z adaptacyjną zmianą współczynników uczenia i momentum. Nauka odbywała się do momentu osiągnięcia przez SSN dopuszczalnego błędu wynoszącego 0,05.

3.4. Wyniki badań

Najlepsze wyniki zostały uzyskane przy wykorzystaniu sieci neuronowej składającej się z 17 neuronów w warstwie wejściowej, dwóch 34 neuronowych warstw ukrytych oraz 6 neuronów wyjściowych. Najskuteczniejszą funkcją okna okazała się być funkcja Dolpha-Czebyszewa, dla której skuteczność identyfikacji stanów emocjonalnych sięgnęła niemal 88% w przypadku rozpoznawania radości. Otrzymane wyniki badań dla poszczególnych funkcji okien czasowych przedstawiono w tabeli 1.

Tabela 1. Wyniki otrzymane dla poszczególnych funkcji okna i stanów emocjonalnych (w procentach)

Funkcja okna	Rodzaj emocji					
	Złość	Znudzenie	Strach	Radość	Neutralny	Smutek
Hamminga	79,17	79,17	80,21	84,38	73,96	76,04
Gaussa	75,00	77,08	71,88	78,13	71,88	73,96
Dolpha–Czebyszewa	**83,33**	**82,29**	**84,38**	**87,50**	**78,13**	**82,29**
Blackmana	78,13	79,17	77,08	82,29	76,04	77,08
Nuttalla	82,29	80,21	81,25	83,33	75,00	79,17
Blackmana-Harrisa	82,29	80,21	82,29	84,38	77,08	80,17

Jak łatwo zauważyć najlepszą skuteczność, bez względu na zastosowane okno, uzyskano dla radości, najgorszą dla stanu neutralnego. Może to być związane z częstotliwością podstawową, która dla radości zdecydowanie różni się od pozostałych stanów emocjonalnych [24]. Średnia wartość amplitudy dla poszczególnych emocji [18] również mogła mieć wpływ na otrzymane wyniki. Należy zauważyć (tabela 2), iż radość była najrzadziej mylonym stanem emocjonalnym, z kolei stan neutralny często był mylony ze znudzeniem oraz smutkiem.

Tabela 2. Macierz pomyłek (w procentach)

Emocje zadane na wejściu	Emocje rozpoznane					
	Złość	Znudzenie	Strach	Radość	Neutralny	Smutek
Złość	**80,04**	3,11	5,03	4,75	3,32	3,75
Znudzenie	1,02	**79,69**	0,94	1,00	9,23	8,12
Strach	7,03	2,13	**79,52**	5,55	3,89	1,88
Radość	5,12	2,98	3,64	**83,34**	2,25	2,67
Neutralny	4,42	8,98	2,22	0,80	**75,35**	8,23
Smutek	2,65	8,11	1,58	1,65	7,89	**78,12**

4. Wnioski

Jak pokazały przeprowadzone badania, rozpoznawanie emocji w sygnale mowy jest zagadnieniem stosunkowo trudnym. Brakuje publikacji bezpośrednio związanych z możliwościami jakie dają metody spektrograficzne oraz sztuczne sieci neuronowe w zagadnieniach związanych z identyfikacją emocji w mowie polskiej.

Podział spektrogramu oraz określenie w jego podobszarach sumarycznej energii pozwoliło na określenie odpowiedniego wektora wejściowego dla SSN.

Przeprowadzone badania pokazały, iż najlepsze wyniki są otrzymywane przy zastosowaniu okna Dolpha-Czebyszewa w procesie wykonywania spektrogramu. Może to być związane z kształtem okna oraz jego skutecznością eliminowania przecieku danych. Warto zauważyć, iż okno Dolpha-Czebyszewa jest efektem optymalizacji, w której to ograniczona została wysokość listków bocznych przy jednoczesnej minimalizacji szerokości listka głównego, co nie ma miejsca w przypadku pozostałych okien poddanych analizie.

Średnia skuteczność SSN przetwarzającej tego rodzaju dane wyniosła około 83%, zaś same wyniki wahają się od około 78% do niemal 88%.

Przeprowadzone badania oraz otrzymane wyniki, jak również wstępne rezultaty kolejnych eksperymentów pozwalają domniemywać, iż zaproponowana metoda przetwarzania sygnału emocjonalnej mowy polskiej może być na tyle uniwersalna, iż będzie możliwe jednoznacznie określenie stanu emocjonalnego mówcy bez względu na rodzaj wypowiedzi.

W dalszym etapie badań planowane jest sprawdzenie możliwości transformaty falkowej i skalogramów w przetwarzaniu emocjonalnej mowy polskiej. Wydaje się, że pozwoli to na opracowanie jeszcze skuteczniejszych metod identyfikacji emocji.

Literatura

[1] Berlin Database of Emotional Speech: http://www.expressive-speech.net/ (10.08.2014).
[2] Bracewell R.: The Fourier Transform and its Application. Electric Engineering Series. McGraw-Hill International Editions. Singapore 2000.
[3] Chena K.F., Lib Y.F.: Combining the Hanning windowed interpolated FFT in both directions. Computer Phisics Communication 178(12)/2008, 924–928.
[4] Chmaj T., Lankosz M.: Akwizycja i przetwarzanie sygnałów cyfrowych. Politechnika Krakowska, Kraków 2011.
[5] Database of Polish Emotional Speech: http://www.eletel.p.lodz.pl/bronakowski/med_cat-alog/ (10.08.2014).
[6] Gałka J., Ziółko B.: Study of Performance Evaluation Methods for Non-Uniform Speech Segmentation, International of Circuits. Systems and Signal Processing. NAUN 2008.
[7] Harris R, Fredric J.: On the use of Windows for Harmonic Analysis with the Discrete Fourier Transform. Proceedings of the IEEE 66(1)/1978, 51–83.
[8] Heinzel, G., Rüdiger, A., Schilling R.: Spectrum and spectral density estimation by the Discrete Fourier transform (DFT), including a comprehensive list of window functions and some new flat-top windows (Technical report). Max Planck Institute (MPI) für Gravitationsphysik/Laser Interferometry & Gravitational Wave Astronomy.
[9] Janicki A., Turkot M.: Rozpoznawanie stanu emocjonalnego mówcy z wykorzystaniem maszyny wektorów wspierających. KSTiT 2008, Bydgoszcz 2008.
[10] Kamińska D., Pelikant A.: Zastosowanie multimedialnej klasyfikacji w rozpoznawaniu stanów emocjonalnych na podstawie mowy spontanicznej. IAPGOŚ 3/2012, 36–39.
[11] Kim E.H., Hyu K.H., Kim S.H., Kwak Y.K.: Speech emotion recognition using eigen-FFT in clean and noisy environments. 16th IEEE International Conference on Robots and Human Interactive Communication, Jeju, Korea 2007.
[12] Kłosiński R.: Materiały X Konferencji Naukowej SP 2014.
[13] Konratowski E.: Czasowo-częstotliwościowa analiza drgań z wykorzystaniem metody overlapping. Logistyka 3/2014, 3104–3110.
[14] Konratowski E.: Monitoring of the Multichannel Audio Signal, Computional collective intelligence. Technologies and Applications. Lecture Notes in Artifical Intelligence 6422, Springer Verlag, 298–306.
[15] Krzyk P., Sułowicz M., Pragłowska–Ryłko N.: Zastosowanie IpDFT do diagnostyki silników asynchronicznych. Zeszyty Problemowe – Maszyny Elektryczne 3/2014, 293–300.
[16] Lynch P.: The Dolph-Chebyshev window: A simple optimal filter. America Meteorological Society Journal of the Online 125/1997, 655–660.
[17] Parsomphan S.: Use of Neural Network Classifier for Detecting Human Emotions via Speech Spectrogram. Procedings of the 3rd IIAE International Conference on Intelligence Systems and Image Processing. Japan 2015.
[18] Pfitzinger H.R., Kaernbach C.: Amplitude and Amplitude Variation of Emotional Speech. Interspeech 2008, 1036–1039.
[19] Powroźnik P., Czerwiński D: Effectiveness comparison on an artificial neural networks to identify Polish emotional speech. Przegląd Elektrotechniczny 07/2016, 45–48.
[20] Powroźnik P.: Polish emotional speech recognition using artificial neural network. Advances is Science and Technology Research Journal 8(24)/2014, 24–27.
[21] Ramakrishnan S.: Recognition of emotion from speech, A review. Speech Enhancement, Modeling and Recognition – Algorithms and Applications, March 2012.
[22] Scherer K.: Vocal communication of emotions: A Review of Research Paradigms in Speech Communication 40/2003, 227–256.
[23] Smith J. O.: Spectral Audio Signal Processing. W3K Publishing, 2011.
[24] Thompson W. F., Balkwill L–L.: Decoding speech prosody in five languages. Semiotica 158/2006, 407–424.
[25] Wicher A., Sęk A., Konieczny J.: Akustyczno-fonetyczne cechy mowy polskiej. Instytut Akustyki UAM Poznań, 2005.
[26] Zieliński T. P., Cyfrowe przetwarzanie sygnałów. Od teorii do zastosowań. WKiŁ, Warszawa 2009.

Mgr Paweł Powroźnik
e-mail: pawel.powroznik@pollub.edu.pl

Absolwent Uniwersytetu Marii Curie Skłodowskiej w Lublinie na kierunki informatyka. Obecnie doktorant w Instytucie Informatyki Politechniki Lubelskiej. Działalność naukowa obejmuje między innymi przetwarzanie sygnału mowy oraz zastosowania sztucznych sieci neuronowych.

Dr hab. inż. Dariusz Czerwiński, prof. PL
e-mail: d.czerwiński@pollub.pl

Absolwent Wydziału Elektrycznego Politechniki Lubelskiej oraz student Uniwersytetu Kanazwa w Japonii. Pracę doktorską obronił w 2001, a habilitacyjną w 2014. Obecnie pełni funkcję dyrektora Instytutu Informatyki Politechniki Lubelskiej.
Działalność naukowa obejmuje między innymi modelowanie numeryczne urządzeń elektromagnetycznych, zastosowania chmur komputerowych, bezpieczeństwo danych w systemach sieciowych, systemy eksploracji danych wykorzystujące paradygmat MapReduce, zastosowania Sztucznych Sieci Neuronowych.

otrzymano/received: 15.06.2016 *przyjęto do druku/accepted: 22.11.2017*